CHAPTER 1

Introduction

A social network is a social structure made up of a set of actors (such as individuals or organizations) and a complex set of the dyadic ties between these actors. Social network mining aims to provide a comprehensive understanding of global and local patterns, mechanism of the network formation, and dynamics of user behaviors. Social network analysis and mining is an inherently interdisciplinary academic field which emerged from sociology, psychology, statistics, and graph theory. However, due to the lack of efficient computational models and the nonavailability of large-scale social networking data, traditional research on social networks has mainly focused on qualitative study in small-scale networks. For example, Milgram spent many years validateng the existence of small-world phenomenon, also referred to as six-degrees of separation by sending mail to thousands of people [107]. In the 1910s, Georg Simmel proposed the concept of structural theories in sociology, which focuses on the dynamic formation of triads [132] and in the 1930s, Jacob Moreno was the first to develop sociograms to analyze people's inter-relationships. Later, for quantitatively analyzing social networks, researchers gave mathematical formulations for social network analysis and developed computational methods [157]. However, most social network research was still conducted by interviewing the participants with small-scale social networks.

More recently, with the emergence and rapid proliferation of online social applications and media—such as instant messaging (e.g., IRC, AIM, MSN, Jabber, Skype), sharing sites (e.g., Flickr, Picassa, YouTube, Plaxo), blogs (e.g., Blogger, WordPress, LiveJournal), wikis (e.g., Wikipedia, PBWiki), microblogs (e.g., Twitter, Jaiku, Weibo), social networks (e.g., Facebook, MySpace, Ning), collaboration networks (e.g., DBLP, ArnetMiner), to mention a few—these online services bring many opportunities for studying social networks, while also posing a number of new challenges. First, the online social network is much larger than the physical social networks in traditional research. Facebook had more than 1.3 billion users in 2014 and Tencent, the largest social networking service in China, attracted 800 million users in 2014. Twitter, the largest microblogging service, has hit half a billion tweets a day. All these big numbers bring new challenges and also the requirements for developing new methods to store, search, analyze, and mine the "big social data." Second, from a microscopic viewpoint, what are the subtle changes of behaviors when users go online from the offline physical networks? Are the traditional social theories still applicable?

Existing research related to social network mining can be categorized into micro-level, meso-level, and macro-level, although the levels of mining are not necessarily mutually exclu-

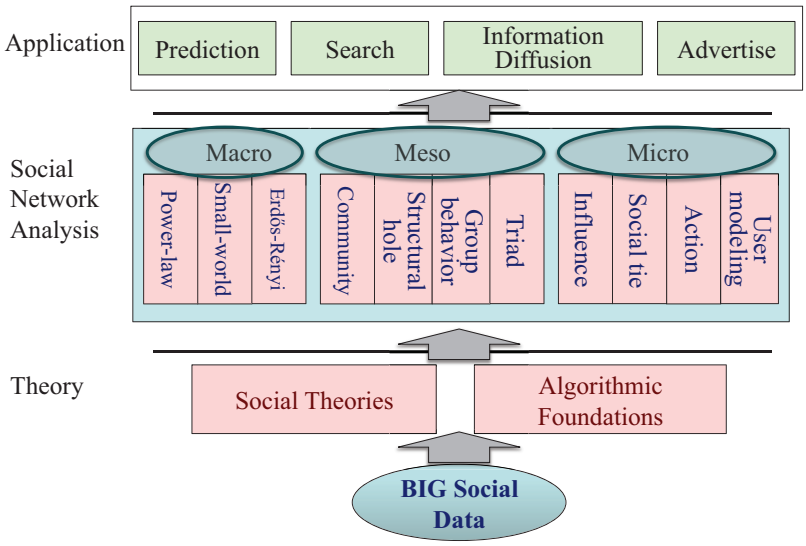

Figure 1.1: Research roadmap for social network mining.

sive. At the micro-level, social network mining is mainly concerned with modeling individuals' behaviors and interactions between users; at the macro-level, it focuses on studying the global patterns of the social networks, for example, network topology [47] and network generative models [10, 46, 158]; and at the micro-level, social network mining is mainly concerned with modeling individuals' behaviors and interactions between users, for example social action theory [159], social ties [56], social influence [79, 143], and user modeling [48]. The meso-level research falls between the micro- and macro-levels, for example community detection [53, 113, 114], structural hole [21, 100], and group behavior analysis [4, 34]. Figure 1.1 gives the research roadmap for social network mining. On top of social network mining, we can consider many applications such as social prediction [138], social search [41, 148], information diffusion [59], and social advertisement [9]. The underlying theories for social network mining include theories from social science and algorithmic foundations from computer science.

In this book, we study social network mining from a microscopic viewpoint. We focus on modeling users' behaviors and their interactions between each other. In particular, we present key technologies for social tie analysis, social influence analysis, and user behavior modeling.

1.1 BACKGROUND

We introduce several related social theories and then briefly review existing literatures on social tie analysis, social influence analysis, and user behavior modeling.

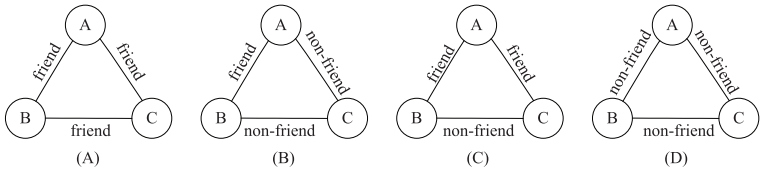

Figure 1.2: Illustration of structural balance theory. (A) and (B) are balanced, while (C) and (D) are not balanced.

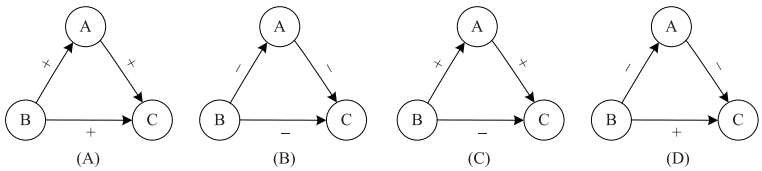

Figure 1.3: Illustration of status theory. (A) and (B) satisfy the status theory, while (C) and (D) do not satisfy the status theory. Here positive "+" denotes the target node has a higher status than the source node; and negative "-" denotes the target node has a lower status than the source node. In total there are 16 different cases.

1.1.1 SOCIAL THEORIES

A basic principle for mining social networks in this book is to incorporate social theories into data mining (or machine learning) model. For social theories, we mainly consider social balance [45], social status [34], structural hole [21], two-step information-flow [89], and strong/weak tie hypothesis [56, 83].

Social balance theory suggests that people in a social network tend to form into a balanced network structure. Figure 1.2 shows such an example to illustrate the structural balance theory over triads, which is the simplest group structure to which balance theory applies. For a triad, the balance theory implies that either all three of these users are friends—"the friend of my friend is my friend"— or only one pair of them are friends—"the enemy of my enemy is my friend."

Another social psychological theory is the *theory of status* [34, 60, 93]. This theory is based on the directed relationship network. Suppose each directed relationship is labeled by a positive sign "+" or a negative sign "-" (where sign "+"/"-" denotes the target node has a higher/lower status than the source node). Then status theory posits that if, in a triangle on three nodes (also-called triad), we take each negative edge, reverse its direction, and flip its sign to positive, then the resulting triangle (with all positive edge signs) should be acyclic. Figure 1.3 illustrates four examples. The first two triangles satisfy the status ordering and the latter two do not satisfy it.

Roughly speaking, a user is said to span a *structural hole* in a social network if she is linked to people in parts of the network that are otherwise not well connected to one another [21]. Such a user is also referred to as *structural hole spanner* [100]. Arguments based on structural holes suggest that there is an informational advantage to have friends in a network who do not know

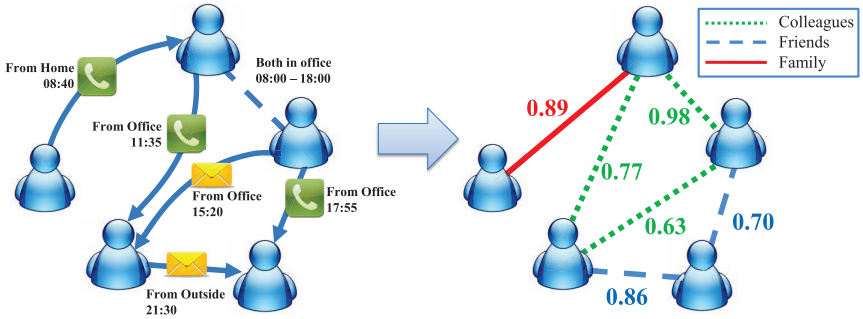

Figure 1.4: An example of inferring social ties in a mobile communication network. The left figure is the input of the task, and the right figure is the output of the task of inferring social ties.

each other. A sales manager with a diverse range of connections can be considered as a structural hole spanner, with a number of potentially *weak ties* [56] to individuals in different communities. More generally, we can think about websites, such as eBay, as spanning structural holes, in that they facilitate economic interactions between people who would otherwise not be able to find each other.

The *two-step information-flow theory* is first introduced in [89] and further elaborated in literature [75, 76]. The theory suggests that ideas (innovations) usually flow first to *opinion leaders*, and then from them to a wider population. In the enterprise email network, for example, managers may act as opinion leaders to help spread information to subordinates.

Interpersonal ties generally come in three varieties: strong, weak, or absent. *Strong tie hypothesis* implies that one's close friends tend to move in the same circles that she/he does, while *weak tie hypothesis* argues that weak ties are responsible for the majority of the embeddedness and structure of social networks in society as well as the transmission of information through these networks [56].

1.1.2 SOCIAL TIE ANALYSIS

Mining social ties is an important problem in social network analysis. Based on the strong/weak tie hypothesis, there is a bunch of research conducted in recent years. The goal of *social tie analysis* is to automatically recognize the semantics associated with each social relationship. Awareness of the semantics of social relationships can benefit many applications. For example, if we could have extracted friendships between users from the mobile communication network, then we can leverage the friendships for a "word-of-mouth" promotion of a new product. Figure 1.4 gives an example of relationship mining in a mobile calling network. The left figure is the input of the problem: in a mobile calling network consisting of users, calls and messages between users, and users' location logs, etc. The objective is to infer the type of the relationships in the network. In the right figure, the users who are family members are connected with red-colored lines, friends are

connected with blue-colored dash lines, and colleagues are connected with green-colored dotted lines. The probability associated with each relationship represents our confidence on the detected relationship types.

There are several works on mining the relationship semantics. Diehl et al. [38] tried to identify the manager-subordinate relationships by learning a ranking function. They defined a ranking objective function and cast the relationship identification as a relationship ranking problem. Menon et al. [106] proposed a log-linear matrix model for dyadic prediction. They used matrix factorization to derive latent features and incorporate the latent features for predicting the label of user relationships. Wang et al. [155] proposed a probabilistic model for mining the advisor-advisee relationships from the publication network. The proposed model is referred to as time-constrained probabilistic factor graph model (TFGM), which supports both supervised and unsupervised learning. Eagle et al. [44] presented several patterns discovered in mobile phone data, and try to use these pattern to infer the friendship network. Tang et al. [149] developed a classification framework of social media based on differentiating different types of social connections. However, these algorithms mainly focus on a specific domain, while our model is general and can be applied to different domains. Moreover, these methods also do not explicitly consider the correlation information between different relationships.

Another research branch is to predict and recommend unknown links in social networks. Liben-Nowell et al. [96] studied the problem of inferring new interactions among users given a snapshot of a social network. They developed several unsupervised approaches to deal with this problem based on measures for analyzing the "proximity" of nodes in a network. The principle is mainly based on similarity of either content or structure between users. Backstrom et al. [6] proposed a supervised random walk algorithm to estimate the strength of social links. Leskovec et al. [92] employed a logistic regression model to predict positive and negative links in online social networks, where the positive links indicate the relationships such as friendship, while negative indicates opposition. However, these works consider only the black-white social networks, and do not consider the types of the relationships.

Recently, Hopcroft et al. [70] explored the problem of reciprocal relationship prediction and Lou et al. [101] extended to study how social relationships develop into triadic closure. They proposed a learning framework to formulate the problem of reciprocal relationship prediction into a graphical model and evaluate the proposed method on a Twitter data set. The framework is demonstrated to be very effective, i.e., it is possible to accurately infer 90% of reciprocal relationships in a dynamic network. Tang et al. [142] further proposed a general framework for classifying the type of social relationships by learning across heterogeneous networks. The idea is to use social theories (e.g., social balance theory, social status theory, structural hole theory, two-step flow theory, and strong/weak tie) as bridge to connect different social networks. Social theory-based features are defined and incorporated into a triad-based factor graph model to infer the type of social relationships in different networks.

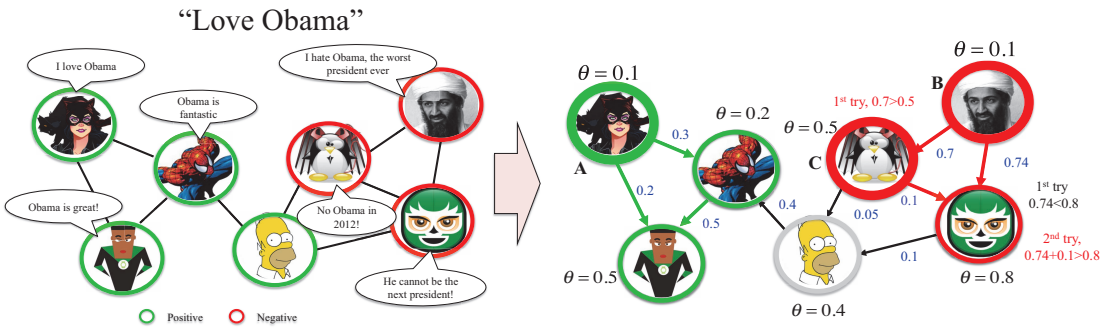

Figure 1.5: An example of social influence for political mobilization. The left figure is the input of the task, and the right figure is the output: influence probability between users, individual conformity of each user, and key influencers (A, B, C).

Another related research topic is relational learning [23, 51]. However, the problem presented in this book is very different. Relational learning focuses on the classification problems when objects or entities are presented in relations, while social tie analysis focuses on exploring the relationship types in social network. A number of supervised methods for link prediction in relational data have been developed [118, 151].

1.1.3 SOCIAL INFLUENCE ANALYSIS

Social influence occurs when one's opinions, emotions, or behaviors are affected by others, intentionally or unintentionally [78]. Recently, social influence analysis has attracted a lot of interest from both research and industry communities. In general, existing research on social influence analysis can be classified into three categories: influence test, influence measure, and influence diffusion models. Figure 1.5 shows an example of social influence for political mobilization. The left figure is the input of the task: opinion of each user for "Obama" in the social network, and the right figure is the output: influence probability between users on this topic "Obama," individual conformity of each user, and key influencers (A, B, C).

Influence Test. Several efforts have been made for identifying the existence of the social influence in the online social networks. For example, Anagnostopoulos et al. [2] gave a theoretical justification to identify influence as a source of social correlation when the time series of user actions is available. They propose a shuffle test to prove the existence of the social influence. Singla and Richardson [133] studied the correlation between personal behaviors and their interests. They found that in the online system people who chat with each other (using instant messaging) are more likely to share interests (their Web searches are the same or topically similar), and the more time they spend talking, the stronger this relationship is. Bond et al. [18] reported results from a randomized controlled trial of political mobilization messages delivered to 61 million Facebook users. They found that when one is aware that their friends have made the political votes, their

likelihood to vote will significantly increase. Crandall et al. [31] further investigated the correlation between social similarity and influence. More recently, some efforts have been made for analyzing the dynamics in the social network. For example, Scripps et al. [128] investigated how different pre-processing decisions and different network forces such as selection and influence affect the modeling of dynamic networks. Other similar work can be referred to Dourisboure et al. [43].

Influence Measure. The goal of influence measure is to quantify the strength of influence between users. Tang et al. [143] introduced the problem of topic-based social influence analysis. They proposed a Topical Affinity Propagation (TAP) approach to describe the problem via using a graphical probabilistic model. However, these works neither consider heterogeneous information nor learn topics and influence strength jointly. Goyal et al. [55] and Saito et al. [124] measured the pairwise influence between two individuals based on the idea of independent cascade model [79]. Liu et al. [99] also studied the problem of measuring the influence on different topics. They proposed a generative graphical model which leverages both heterogeneous link information and textual content associated with each user in the network to mine topic-level influence strength. Based on the learned direct influence, we further study the influence propagation and aggregation mechanisms: conservative and non-conservative propagations to derive the indirect influence. Xin et al. [131] studied the indirect influence using the theory of quantum cognition. Myers et al. [112] proposed a probabilistic model to quantify the external influence out-of-network sources. Belak et al. [12] investigated and measured the influence on the cross-community level so as to provide a coarse-grained picture of a potentially very large network. They presented a framework for cross-community influence analysis and evaluated the proposed method on a ten-year data set from the largest Irish online discussion system Boards.ie, Boards.ie is the name of the system. Zhang et al. [174] proposed the notion of social influence locality and applied it for modeling users' retweeting behaviors in the social networks. They developed two instantiation functions based on pairwise influence and structural diversity.

Influence Diffusion Models. Social influence has been applied in the application of influence maximization in viral marketing. Domingos and Richardson [42, 121] were the first to study influence maximization as an algorithmic problem. Kempe et al. [79] took the first step to formulating influence maximization as a discrete optimization problem. Leskovec et al. [94] and Chen et al. [26, 27] made efforts to improve the efficiency of influence maximization. Gruhl et al. [59] proposed a time-decayed diffusion model for blogging writing, and use an EM-like algorithm to estimate the influence probabilities. Yang et al. [167] studied the interplay between users' social roles and their influence on information diffusion. They proposed a Role-Aware INformation diffusion model (RAIN) that integrates social role recognition and diffusion modeling into a unified framework.

1.1.4 USER MODELING AND ACTIONS

User modeling describes the process of building up a user model to characterize users' skills, declarative knowledge, and specific needs to a system [48].

A number of models have been proposed to model users' behaviors in dynamic social networks. Sarkar et al. [125] developed a generalized model associating each entity in Euclidean latent space and used kernel functions for similarity in latent space to model friendship drifting over time. Tan et al. [138] studied how users' behaviors (actions) in a social network are influenced by various factors such as personal interests, social influence, and global trends. They proposed a Noise Tolerant Time-varying Factor Graph Model (NTT-FGM) for modeling and predicting social actions, which simultaneously models social network structure, user attributes, and user action history for better prediction of the users' future actions. Tan et al. [137] investigated how users' sentiment can be inferred in the social network by incorporating the social network information. Scripps et al. [128] presented a model to investigate how different pre-processing decisions and different network forces such as selection and influence affect the modeling of dynamic networks. They also demonstrated the effects of attribute drifting and the importance of individual attributes in forming links over time.

Group analysis is based on the view that deep lasting change can occur within a carefully formed group whose combined membership reflects the wider norms of society. There is an interest, in group analysis, on the relationship between the individual group member and the rest of the group resulting in a strengthening of both, and a better integration of the individual with his or her community, family and social network. Shi et al. [130] studied the pattern of user participation behavior, and the feature factors that influence such behavior on different forum data sets. Backstrom et al. [5] proposed a partitioning on the data that selects for active communities of engaged individuals.

1.1.5 GRAPHICAL MODELS

From the perspective of machine learning and data mining, for analyzing networking data, graphical probabilistic models are often employed to describe the dependencies between observation data. Bayesian networks [72], Markov random field [134], factor graph [50, 84], Restricted Boltzmann Machine (RBM) [162], deep neural networks [68], and many others are widely used graphical models.

1.2 BOOK OUTLINE

Despite much work having been conducted for mining social networks, we are still facing a number of challenges. In this book, from the microscopic viewpoint, we introduce key technologies for social tie analysis, social influence analysis, and user behavior modeling. Figure 1.6 gives an overview of the topics covered in this book and their relationships. Specifically, the book is organized as follows.

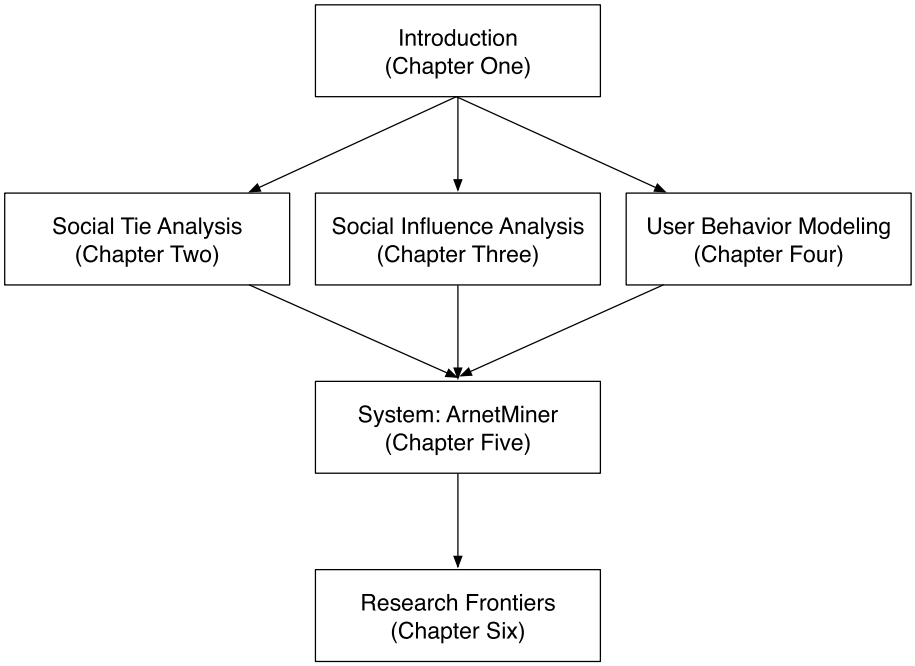

Figure 1.6: Outline of this book.

- In Chapter 2, we begin with the analytics of social interactions between users. The first question we try to answer is: "What are the fundamental factors that form the different categories of social ties?." We incorporate social network theories into a probabilistic graphical model for learning to automatically infer the types of social relationships.
- In Chapter 3, we investigate the problem of social influence. We introduce theoretical methodologies for verifying the existence of social influence, and for quantifying the influence strength between users on different topics.
- In Chapter 4, we study how to model and predict user behaviors. One fundamental problem is to distinguish the effects of different social factors such as social influence, homophily, and individual's characteristics. We introduce a probabilistic model to address this problem.
- In Chapter 5, we use ArnetMiner, an academic social network developed by the authors, as an example to demonstrate how we apply the introduced technologies for mining real social networks at the micro-level.
- Finally, in Chapter 6, we present potential future directions in this field.